Upper limits from counting experiments with multiple pipelines

Patrick J Sutton

School of Physics and Astronomy, Cardiff University, Cardiff, United Kingdom, CF24 3AA

E-mail: patrick.sutton@astro.cf.ac.uk

Abstract. In counting experiments, one can set an upper limit on the rate of a Poisson process based on a count of the number of events observed due to the process. In some experiments, one makes several counts of the number of events, using different instruments, different event detection algorithms, or observations over multiple time intervals. We demonstrate how to generalize the classical frequentist upper limit calculation to the case where multiple counts of events are made over one or more time intervals using several (not necessarily independent) procedures. We show how different choices of the rank ordering of possible outcomes in the space of counts correspond to applying different levels of significance to the various measurements. We propose an ordering that is matched to the sensitivity of the different measurement procedures and show that in typical cases it gives stronger upper limits than other choices. As an example, we show how this method can be applied to searches for gravitational-wave bursts, where multiple burst-detection algorithms analyse the same data set, and demonstrate how a single combined upper limit can be set on the gravitational-wave burst rate.

PACS numbers: 06.20.Dk, 04.80.Nn

1. Introduction

One of the most familiar applications of classical confidence intervals is to the counting experiment, in which one attempts to measure or place a limit on the rate of a physical Poisson process by counting the number of occurrences of the process observed during some period of time. For example, for a single measurement (a single count of events) with low background and an expected physical rate comparable to or lower than the background, one typically sets an upper limit; *i.e.*, a one-sided confidence interval. Given a count n, the upper limit is that value of the physical rate such that the *a priori* probability of measuring more than n events in the experiment exceeds some chosen confidence level.

Various issues may complicate the procedure for setting the upper limit. For example, if the background is large, there is a well-known problem that the upper confidence limit may be the empty set when the observed number of events is much lower than that expected from the background. Another more subtle issue is that the decision to report an upper limit versus a two-sided confidence interval can, if based on the data, cause undercoverage, rendering the procedure invalid. Techniques for addressing these issues have been presented in the literature, for example, by the Feldman-Cousins technique [1] and the loudest event technique [2, 3]. These can also be addressed by Bayesian methods; see for example [3, 4, 5, 6, 7, 8].

In this paper we are concerned with a different complication: when more than one count is made of the number of events. One example of where this situation arises is searches for gravitational-wave (GW) bursts with LIGO and similar detectors [9, 10, 11]. In this scenario the GW signals are expected to have amplitudes near the noise floor of the detectors, and the rate of detectable events is expected to be of order the inverse of the observation time or less. To improve chances of detection, multiple algorithms are used to analyse the data [12, 13, 14], each producing its own list of candidate GW bursts. The event lists produced by these algorithms, however, are not completely independent. They will generally show some correlation between which foreground events they detect, and may also show some correlation between the background noise fluctuations they detect. Furthermore, the data set itself typically is not of uniform sensitivity. For example, the longest data-collection run to date for the LIGO-GEO-Virgo network lasted more than two years [15]. Over this time the sensitivity of each of the instruments changed, and at any given time during the run, anywhere between 1 and 5 detectors may have been operating. The challenge to the data analyst in such an experiment is this: given multiple counts of events collected from processing several data sets of different sensitivities and with different algorithms, how does one set a single limit on the physical event rate?

There are many options. The simplest is to take the union of all of the event lists and observation time, effectively converting the multiple observations into a single observation, and computing the upper limit using a standard technique. This approach ignores differences in the quality of the data from the different epochs, and in the algorithms themselves. Alternatives include discarding results from select data sets or algorithms (presumably the less sensitive ones), again with the aim of reducing the observations to effectively a single count. These approaches invariably involve loss of information from the experiment. Intuitively, one expects to be able to set stronger limits if one uses all of the information from the experiment rather than only a subset of the information.

In this article we propose a general formalism for setting classical upper limits on experiments involving multiple *pipelines*, where a pipeline denotes the analysis of a single data set by a single algorithm. We characterize the observational results and the sensitivities of the experiment in terms of logical combinations of pipelines. We show that various choices such as taking the union of data sets correspond to particular choices of weighting of measurements. We propose a specific weighting choice based on the efficiencies (sensitivities) of the logical combinations, and show that it gives stronger upper limits than other choices in typical cases. Furthermore, the efficiency weighting choice makes use of all of the experiment results, naturally handles correlated measurements, and tends to be robust against occasional background contamination of counts.

This paper is organized as follows. In Section 2 we review how one sets a classical upper limit on the rate of a Poisson-distributed process in a counting experiment. In Section 3 we generalize the single-count procedure to the case of multiple counts. We discuss various choices of the weighting to obtain upper limits, including our sensitivity-based proposal. We demonstrate each procedure for the case of a counting experiment using two pipelines, with and without background. In Section 4 we demonstrate how the same procedure naturally handles multiple data sets. Section 5 contains a few brief remarks on the applicability of the method.

2. Single-Pipeline Case

We briefly review how one sets a classical upper limit (a one-sided confidence interval) on the rate of a Poisson-distributed process via a counting experiment.

Consider an experiment that measures the number of events of a specific random process that occurs in a time T. We assume that the foreground events occur independently of one another, with a mean rate μ that is unknown *a priori*. We further assume that the experiment has a probability ϵ of successfully detecting (counting) any given event. Finally, we assume that the mean number of background events (due to "noise" or effects other than the physical effect of interest) in time T is b. Then the actual total number of events (foreground plus background) that will be counted in a given time T is Poisson distributed, as is easily demonstrated.

Let us divide the observation time T into M equal sub-intervals of length T/M. In the limit of large M, the probability of one event being detected in any given sub-interval is $(\epsilon\mu T + b)/M \ll 1$, and the probability of more than one event in the same interval is negligible. The probability of detecting a total of N events over the full time T is derived from binomial statistics as the probability of N "successes" in M "trials". Defining $\lambda \equiv \mu T$ as the expected mean number of foreground events occurring, we have

$$\begin{aligned} P(N|\epsilon\lambda + b) &= \lim_{M\to\infty} \binom{M}{N} \left(\frac{\epsilon\lambda + b}{M}\right)^{N} \left(1 - \frac{\epsilon\lambda + b}{M}\right)^{M-N} \\ &= \frac{(\epsilon\lambda + b)^N}{N!} \mathrm{e}^{-(\epsilon\lambda+b)} \, . \end{aligned} \tag{1}$$

This is the familiar Poisson distribution for a process with mean number of detected events $\epsilon\lambda + b$.

Given an actual measured number n, the Poisson distribution (1) can be used to set an upper limit on the value of λ, or equivalently on μ. Heuristically, values of λ much larger than $(n - b)/\epsilon$ are unlikely to produce only n detected events. More formally, we select a confidence level $\alpha \in (0, 1)$. The frequentist upper limit λ_α at confidence level α given n measured events is that value of λ at which there is an *a priori* probability α of measuring more than n events. Implicitly, λ_α is given by

$$\alpha = \sum_{N=n+1}^{\infty} P(N|\epsilon\lambda_\alpha + b) = 1 - \sum_{N=0}^{n} P(N|\epsilon\lambda_\alpha + b)\,. \tag{2}$$

We define the cumulative probability $C(n|\epsilon\lambda+b)$ as the *a priori* probability of detecting n or fewer events:

$$C(n|\epsilon\lambda + b) \equiv \sum_{N=0}^{n} P(N|\epsilon\lambda + b)\,. \tag{3}$$

We can write the upper limit formula for λ_α as

$$C(n|\epsilon\lambda_\alpha + b) = 1 - \alpha\,. \tag{4}$$

For example, the 90% confidence level ($\alpha = 0.9$) upper limit for zero observed events ($n = 0$) and zero background ($b = 0$) is

$$0.1 = C(0|\epsilon\lambda_{90\%}) = \mathrm{e}^{-\epsilon\lambda_{90\%}}\,, \quad \lambda_{90\%} = \frac{2.30}{\epsilon}\,. \tag{5}$$

For $n = 1$ observed events the upper limit is higher (weaker):

$$0.1 = (1 + \epsilon\lambda_{90\%})\mathrm{e}^{-\epsilon\lambda_{90\%}}\,, \quad \lambda_{90\%} = \frac{3.89}{\epsilon}\,. \tag{6}$$

To be rigorous, one must prove that the upper limit formula (4) has a coverage of at least α. The coverage is defined as the fraction of measurements in an ensemble of identical experiments for which the derived upper limit is greater than or equal to the true rate λ_{true}. To be a valid upper limit with confidence level α, one must show that $\lambda_\alpha \geq \lambda_{\mathrm{true}}$ in a fraction $\geq \alpha$ of experiments *for any possible value of* λ_{true}.

It is straightforward to prove that the upper limit formula (4) has the coverage α. First, we note two properties ‡ of $C(n|\epsilon\lambda + b)$:

$$C(n|\epsilon\lambda + b) \quad > C(m|\epsilon\lambda + b) \qquad \text{for } n > m\,; \tag{7}$$

$$\frac{\mathrm{d}C(n|\epsilon\lambda + b)}{\mathrm{d}\lambda} < 0\,. \tag{8}$$

Let us suppose that the true value of the rate is λ_{true}. Let m be the largest integer such that $C(m|\epsilon\lambda_{\mathrm{true}} + b) \leq 1 - \alpha$. By definition of m, in a fraction $\geq \alpha$ of experiments the measured number of events n will be larger than m. For these cases $C(n|\epsilon\lambda_{\mathrm{true}} + b) > 1 - \alpha$. Applying the upper limit formula (4) and noting (8), we see that in these cases the derived upper limit λ_α will be greater than λ_{true}. The coverage is thus established.

We should note that one has the freedom to ignore the experimental background when computing the upper limit; *i.e.*, one may use the approximation $b = 0$. Since the background will increase n above the value due to the physics of interest, the upper limit derived using $b = 0$ remains valid (provides minimum coverage), though it will be higher than if we had accounted for the background. We will use this approximation in some of our worked examples.

‡ From (1), $\mathrm{d}C(n|\epsilon\lambda + b)/\mathrm{d}\lambda = -\epsilon(\epsilon\lambda + b)^n \mathrm{e}^{-\epsilon\lambda - b}/n! < 0$ for $\lambda > 0$, $b \geq 0$.

We also note the well-known phenomenon that the classical one-sided confidence interval procedure can produce an empty upper limit when the number of observed events is much lower than the background. For example, for $n = 0$ observed events and $b = 3$ the 90% upper limit is the solution of

$$0.1 = C(0|\epsilon\lambda_{90\%} + 3) = \mathrm{e}^{-\epsilon\lambda_{90\%}-3}\,. \tag{9}$$

This has no solution with $\lambda_{90\%} \geq 0$. Methods for handling this issue have been proposed, for example, by Feldman and Cousins [1]. In this paper we consider only one-sided confidence intervals, and therefore we will restrict ourselves to case where $b \lesssim 1$.

3. Multiple-Pipeline Case

3.1. Formulation

The simplest example of a multiple-pipeline experiment is one in which two different methods or "pipelines" are used to count events (by processing the same data, watching the same sky, *etc.*) over the same epoch T. (We'll consider the case of *disjoint* data sets in Section 4.) Denote the pipelines by A and B. Any given event may be detected by pipeline A only, by pipeline B only, by both A and B, or by neither pipeline. We characterize the sensitivity of the experiment by the three numbers ϵ_A, ϵ_B, and ϵ_{AB}:

ϵ_A: The probability that any given foreground event will be detected by pipeline A but not detected by pipeline B;

ϵ_B: The probability that any given foreground event will be detected by pipeline B but not detected by pipeline A;

ϵ_{AB}: The probability that any given foreground event will be detected by both pipeline A and pipeline B.

We denote the expected background by the three numbers b_A, b_B, and b_{AB}:

b_A: The expected number of background events detected by pipeline A but not detected by pipeline B;

b_B: The expected number of background events detected by pipeline B but not detected by pipeline A; and

b_{AB}: The expected number of background events detected by both pipeline A and pipeline B.

Finally, the outcome of the counting experiment is the set of three numbers n_A, n_B, and n_{AB}:

n_A: The number of events detected by pipeline A but not detected by pipeline B;

n_B: The number of events detected by pipeline B but not detected by pipeline A; and

n_{AB}: The number of events detected by both pipeline A and pipeline B.

To interpret (n_A, n_B, n_{AB}) in terms of an upper limit on λ, we first need to compute the joint probability $P((n_A, n_B, n_{AB}|\lambda, \epsilon_A, \epsilon_B, \epsilon_{AB}, b_A, b_B, b_{AB})$. This is straightforward; repeating the logic of the single-pipeline case, it is easy to see that

$$\begin{aligned}
&P(N_A, N_B, N_{AB}|\lambda, \epsilon_A, \epsilon_B, \epsilon_{AB}, b_A, b_B, b_{AB}) \\
&= \lim_{M\to\infty} \binom{M}{N_A} \binom{M-N_A}{N_B} \binom{M-N_A-N_B}{N_{AB}} \\
&\quad \times \left(\frac{\epsilon_A\lambda + b_A}{M}\right)^{N_A} \left(\frac{\epsilon_B\lambda + b_B}{M}\right)^{N_B} \\
&\quad \times \left(\frac{\epsilon_{AB}\lambda + b_{AB}}{M}\right)^{N_{AB}} \left(1 - \frac{\epsilon_{TOT}\lambda + b_{TOT}}{M}\right)^{M-N_{TOT}} \\
&= P(N_A|\epsilon_A\lambda + b_A) P(N_B|\epsilon_B\lambda + b_B) P(N_{AB}|\epsilon_{AB}\lambda + b_{AB}) \,. \qquad (10)
\end{aligned}$$

Here we have defined the total number of events detected,

$$N_{TOT} \equiv N_A + N_B + N_{AB} \,, \qquad (11)$$

the total number of events expected from background,

$$b_{TOT} \equiv b_A + b_B + b_{AB} \,, \qquad (12)$$

and the probability of a given foreground event being detected by any combination of pipelines,

$$\epsilon_{TOT} \equiv \epsilon_A + \epsilon_B + \epsilon_{AB} \,. \qquad (13)$$

We see that by choosing to characterize the outcome of the experiment by the number of events detected by logical combinations of pipelines, the joint probability factorizes to the product of single-pipeline probabilities (1). The measurements of N_A, N_B, and N_{AB} can therefore be regarded as statistically independent experiments. This is a key simplification that makes deriving a combined upper limit straightforward.

In the general case of p pipelines, there are $q \equiv 2^p - 1$ distinct combinations by which an event may be detected. Using the vector notation $\vec{N}$, $\vec{\epsilon}$, and $\vec{b}$, where the vector index $i \in [1, \ldots, q]$ labels the distinct combinations, we have

$$P(\vec{N}|\lambda\vec{\epsilon} + \vec{b}) = \prod_{i=1}^{q} P(N_i|\lambda\epsilon_i + b_i) \,. \qquad (14)$$

3.2. Defining an Upper Limit

To set an upper limit we need first to define a cumulative probability distribution $C(\vec{n}|\lambda\vec{\epsilon} + \vec{b})$ corresponding to (14), analogous to (3). Since the space of observation $\{\vec{N}\}$ is multi-dimensional, we have a great deal of freedom in how we choose to sum over $\{\vec{N}\}$ to define the cumulative distribution. Put another way, we must chose a *rank ordering* of $\{\vec{N}\}$. (For an unbiased limit, this must be done before the measurement of $\vec{n}$.)

To construct a confidence belt, we choose a one-parameter family of surfaces $\mathcal{S}(\zeta)$ that foliates the observation space $\{\vec{N}\}$. This family is chosen so that for every value of the parameter ζ, the surface $\mathcal{S}(\zeta)$ divides the space $\{\vec{N}\}$ into two regions: a *acceptance region* of low number of events (including the origin, and the surface $\mathcal{S}(\zeta)$ itself), and a *rejection region* of high number of events. Our choice of the family $\mathcal{S}(\zeta)$ is arbitrary, except that the outward normal to each surface must have non-negative components everywhere; this is required to prove coverage, as shown below. As we shall see, our freedom in the choice of the $\mathcal{S}(\zeta)$ corresponds to how the various pipelines are "weighted" in contributing to the upper limit.

Because of the foliation, every point $\vec{N}$ in the observation space lies on exactly one such surface, which we refer to as an *exclusion surface*. Hence, each point $\vec{N}$ can be associated with a single parameter value, $\zeta(\vec{N})$. This gives us a rank ordering of the $\vec{N}$ defining whether a given $\vec{N}'$ contains "more," the "same," or "fewer" events than $\vec{N}''$. The family $\mathcal{S}(\zeta)$ therefore maps the multi-dimensional space $\{\vec{N}\}$ to a one-dimensional space. This allows us to define a cumulative probability $C_{\mathcal{S}}(\vec{n}|\lambda\vec{\epsilon}+\vec{b})$ by

$$C_{\mathcal{S}}(\vec{n}|\lambda\vec{\epsilon}+\vec{b}) \equiv \sum_{\vec{N}|\zeta(\vec{N})\leq\zeta(\vec{n})} P(\vec{N}|\lambda\vec{\epsilon}+\vec{b})\,, \tag{15}$$

where the sum is taken over all $\vec{N}$ for which $\zeta(\vec{N}) \leq \zeta(\vec{n})$; *i.e.*, over all $\vec{N}$ that contain as few events or fewer than $\vec{n}$.

Given a family of exclusion surfaces $\mathcal{S}(\zeta)$ and a measured number of events $\vec{n}$, we may use the cumulative probability $C_{\mathcal{S}}$ to set an upper limit on λ in the same way as is done for the single-pipeline case. Specifically, for a measured number of events $\vec{n}$, the upper limit λ_α at confidence level α is

$$C_{\mathcal{S}}(\vec{n}|\lambda_\alpha\vec{\epsilon}+\vec{b}) = 1-\alpha\,. \tag{16}$$

That is, the upper limit λ_α on the rate is that value for which in a fraction α of an ensemble of experiments one would measure a number of events that falls in the rejection region of $\mathcal{S}(\zeta(\vec{n}))$. Put another way, the upper limit is the rate for which one should measure "more" than $\vec{n}$ events (a value of ζ larger than $\zeta(\vec{n})$) in a fraction α of an ensemble of experiments.

We will consider various simple choices of families $\mathcal{S}(\zeta)$ and their interpretations shortly. First, however, we prove that the algorithm (16) has coverage α.

3.3. Coverage

We now prove that the upper limit formula (16) has a coverage of at least α. The proof follows that for the single-pipeline case in Section 2. Again, we note two properties of $C_{\mathcal{S}}(\vec{n}|\lambda\vec{\epsilon}+\vec{b})$:

$$C_{\mathcal{S}}(\vec{n}|\lambda\vec{\epsilon}+\vec{b}) \quad > C_{\mathcal{S}}(\vec{m}|\lambda\vec{\epsilon}+\vec{b}) \ \text{ for } \ \zeta(\vec{n}) > \zeta(\vec{m})\,; \tag{17}$$

$$\frac{\mathrm{d}C_{\mathcal{S}}(\vec{n}|\lambda\vec{\epsilon}+\vec{b})}{\mathrm{d}\lambda} < 0\,. \tag{18}$$

(See Appendix Appendix A for the proof of (18).) Let us suppose that the true value of the rate is λ_{true}. Let m be the vector with nonnegative integer components and with the largest value of $\zeta(\vec{m})$ such that $C_{\mathcal{S}}(\vec{m}|\lambda_{\text{true}}\vec{\epsilon}+\vec{b}) \leq 1-\alpha$. By definition of m, in a fraction $\geq \alpha$ of experiments the measured number of events $\vec{n}$ will have $\zeta(\vec{n}) > \zeta(\vec{m})$. For these cases $C_{\mathcal{S}}(\zeta(\vec{n})|\lambda_{\rm true}\vec{\epsilon} + \vec{b}) > 1 - \alpha$. Applying the upper limit formula (16) and noting (18), we see that in these cases the derived upper limit λ_α will be greater than $\lambda_{\rm true}$. The coverage is thus established.

As stated before, our choice of exclusion surfaces is arbitrary except that the outward normal to the contour must have non-negative components everywhere. This restriction ensures that equation (18) is valid, which in turn is required to prove coverage. As in the single-pipeline case, we may chose to ignore the background and use $\vec{b} = 0$ when computing upper limits. Since a non-zero background contribution will increase the measured ζ over its zero-background value, from (16)-(18) it follows that the limit will be higher than that computed accounting for the background, but coverage will be maintained.

3.4. Choosing Exclusion Surfaces

We now turn to the question of how to select the family of exclusion surfaces to obtain the strongest limits. For simplicity, we restrict ourselves henceforth to the simple case of plane surfaces. In this case, a family of exclusion surfaces is set by choosing the vector $\vec{k}$ that is normal to the planes. The parameter for the family is then $\zeta(\vec{N}) = \vec{k} \cdot \vec{N}$ (the magnitude of $\vec{k}$ is irrelevant). For a given observation $\vec{n}$ the upper limit λ_α is given by

$$C_{\vec{k}}(\vec{n}|\lambda_\alpha\vec{\epsilon} + \vec{b}) \equiv \sum_{\vec{N}|(\vec{n}-\vec{N})\cdot\vec{k}\geq 0} P(\vec{N}|\lambda_\alpha\vec{\epsilon} + \vec{b}) = 1 - \alpha\,. \tag{19}$$

Note that the sum is taken over all $\vec{N}$ satisfying the condition

$$(\vec{n} - \vec{N}) \cdot \vec{k} \geq 0\,. \tag{20}$$

We now explore several simple choices of exclusion surfaces with ready physical interpretations: taking the logical AND or OR combinations of pipelines, and using only the most sensitive pipeline. We then propose a new choice of exclusion surfaces: $\vec{k} = \vec{\epsilon}$; *i.e.*, we weight the measurements by the relative sensitivity of their pipelines. We show that this efficiency-weighted approach has several advantages over the other choices discussed. In particular, it gives upper limits that are better than those from the other common choices for most outcomes of the experiment.

3.4.1. OR combination One obvious way to orient the exclusion surfaces is to set the normal vector $\vec{k} = (1, 1, \ldots, 1)$. This choice treats all distinct pipeline combinations equally. For a given observation $\vec{n}$ the upper limit on λ is then given by (16) with the sum taken over all $\vec{N}$ satisfying the condition

$$\sum_i N_i \leq \sum_i n_i\,, \tag{21}$$

or simply

$$N_{TOT} \leq n_{TOT}\,. \tag{22}$$

That is, the upper limit depends only on the total number of events detected, regardless of which pipelines or combinations of pipelines detected them. We see that this choice of exclusion contour is equivalent to setting an upper limit based on a single pipeline which is formed by taking the "OR" combination of all events detected by all pipelines or combinations of pipelines.

For example, consider the case of two pipelines A and B. Let us assume for simplicity that the background is negligible ($b_A, b_B, b_{AB} \simeq 0$). If no events are detected, the upper limit at confidence level $\alpha = 0.9$ is given by

$$\begin{aligned} 0.1 &= C_{\vec{k}}((0,0,0)|\lambda_{90\%}\vec{\epsilon}) \\ &= \mathrm{e}^{-\epsilon_{TOT}\lambda_{90\%}} \,, \end{aligned} \tag{23}$$

where $\epsilon_{TOT} \equiv \epsilon_A + \epsilon_B + \epsilon_{AB}$. This has the solution

$$\lambda_{90\%} = \frac{2.30}{\epsilon_{TOT}} \,. \tag{24}$$

This has the same form as in the single-pipeline case, (5), with the replacement $\epsilon \to \epsilon_{TOT}$.

Now consider the case of one event detected (it does not matter whether the lone event is detected by A, by B, or by both). The upper limit is given by

$$\begin{aligned} 0.1 &= P((0,0,0)|\lambda_{90\%}\vec{\epsilon}) + P((1,0,0)|\lambda_{90\%}\vec{\epsilon}) \\ &\quad + P((0,1,0)|\lambda_{90\%}\vec{\epsilon}) + P((0,0,1)|\lambda_{90\%}\vec{\epsilon}) \\ &= (1 + \epsilon_{TOT}\lambda_{90\%})\mathrm{e}^{-\epsilon_{TOT}\lambda_{90\%}} \,, \end{aligned} \tag{25}$$

which has the solution

$$\lambda_{90\%} = \frac{3.89}{\epsilon_{TOT}} \,. \tag{26}$$

This again has the same form as in the single-pipeline case, (6), with the replacement $\epsilon \to \epsilon_{TOT}$.

The OR combination has the advantage that it has the largest efficiency of any combination, since an event is counted if any of the pipelines detect it. This leads to strong upper limits when no events are detected. The disadvantage is that the background is also summed over all pipeline combinations, potentially leading to a high false alarm rate and poor limits if any of the pipeline samples are contaminated by background.

3.4.2. AND combination A "conservative," choice for detecting events is to demand that *all* pipelines observe an event for it to be counted as a possible signal. It is easy to see that this is equivalent to choosing contours with normal vector $\vec{k} = (0, \ldots, 0, 1)$. The upper limit for observation $\vec{n}$ is then given by (16) with the sum is taken over all $\vec{N}$ satisfying the condition

$$N_q \leq n_q \,, \tag{27}$$

where n_q is the number of events detected in coincidence by all pipelines. Because of the factorization of the joint probability (14), the upper limit becomes

$$\begin{aligned}
1 - \alpha &= \sum_{N_1=0}^{\infty} \cdots \sum_{N_{q-1}=0}^{\infty} \sum_{N_q=0}^{n_q} P(\vec{N} | \lambda_\alpha \vec{\epsilon} + \vec{b}) \\
&= \left[\sum_{N_1=0}^{\infty} P(N_1 | \epsilon_1 \lambda_\alpha + b_1) \right] \cdots \\
&\quad \times \left[\sum_{N_{q-1}=0}^{\infty} P(N_{q-1} | \epsilon_{q-1} \lambda_\alpha + b_{q-1}) \right] \\
&\quad \times \sum_{N_q=0}^{n_q} P(N_q | \epsilon_q \lambda_\alpha + b_q) \\
&= \sum_{N_q=0}^{n_q} P(N_q | \epsilon_q \lambda_\alpha + b_q) \,, \qquad (28)
\end{aligned}$$

We see that the upper limit reduces to that for an effective single pipeline formed by taking the AND combination of all pipelines. This has the same form as in the single-pipeline case, (5), with the replacement $\epsilon \to \epsilon_q$.

Consider again the case of two pipelines A and B with low background. Suppose we had decided *a priori* to compute an AND upper limit. If no events were detected by any pipeline, then the 90% confidence upper limit is given by (5) with $\epsilon \to \epsilon_{AB}$.

$$\lambda_{90\%} = \frac{2.30}{\epsilon_{AB}} \,. \qquad (29)$$

Since $\epsilon_{AB} \leq \epsilon_{TOT}$, the AND combination gives a weaker limit for a given number of measured events.

Now consider the case in which one event is detected. The limit now depends on which pipeline combination detected the event. If only one of the pipelines detected the event, then $n_q = 0$, and the 90% confidence upper limit is given by (29). If both pipelines detected the event then $n_q = 1$ and

$$\lambda_{90\%} = \frac{3.89}{\epsilon_{AB}} \,. \qquad (30)$$

These have the same form as in the single-pipeline case, (5), (6), with the replacement $\epsilon \to \epsilon_{AB}$.

The AND combination has the advantage of being the combination least susceptible to background contamination, since an event is only counted if it is detected by all pipelines. For example, the AND combination is particularly robust if the pipelines have different responses to the background noise. The disadvantage is that the efficiency is also the lowest of any combination, for the same reason. In particular, the AND sensitivity is limited by the least sensitive pipeline.

3.4.3. SINGLE combination Another simple choice for setting the upper limit is to consider only the measurement by the single most sensitive pipeline, and ignoring all of the others. The most sensitive pipeline is the one with the largest detection efficiency computed when ignoring the other pipelines; *e.g.*, for the two-pipeline case it is the larger of $\epsilon_A + \epsilon_{AB}$ (for A) or $\epsilon_B + \epsilon_{AB}$ (for B). The procedure for computing the upper limit in this case is simply to apply (4). We note here that it is another special case of the multiple-pipeline procedure. For example, for two pipelines where A is the more sensitive, the SINGLE limit is equivalent to choosing

$$k = (1, 0, 1)\,. \tag{31}$$

If no events are detected by A, then the 90% confidence upper limit is given by (5) with $\epsilon \to \epsilon_A + \epsilon_{AB}$:

$$\lambda_{90\%} = \frac{2.30}{\epsilon_A + \epsilon_{AB}}\,. \tag{32}$$

If one event is detected by A, the limit is

$$\lambda_{90\%} = \frac{3.89}{\epsilon_A + \epsilon_{AB}}\,. \tag{33}$$

The efficiency and background of the SINGLE combination are intermediate between those of the OR and AND combinations. In general, $\epsilon_{TOT} \geq \epsilon_A + \epsilon_{AB} \geq \epsilon_{AB}$, so for a given number of measured events (for example, 0), OR will give the strongest limit, AND the weakest, and SINGLE an intermediate value. On the other hand, the background is highest for OR and lowest for AND, so there is a greater chance of having $n > 0$ events in the OR combination. Unfortunately, for an unbiased analysis one must choose the upper limit method *before* counting events, so it is difficult to make the best choice between the AND, OR, and SINGLE options *a priori*.

3.4.4. Efficiency-weighted combination The AND, OR, and SINGLE options are just three examples of how one may select the exclusion surfaces for the multiple-pipeline counting experiment. As just discussed, the relative strength of the upper limits one can achieve with these options depends on the number of events detected by each pipeline combination, which one does not know *a priori* in a blind analysis.

An obvious drawback of the AND and OR examples is that the exclusion surfaces are selected without regard to the known sensitivities $\vec{\epsilon}$ of the various pipeline combinations. One expects that the strongest upper limits should involve use of this information. As a trivial example, a pipeline combination with zero detection probability ($\epsilon_i = 0$) should be ignored when setting upper limits (n_i should be ignored). The SINGLE combination makes some limited use of the known sensitivities, but throws away all of the information produced by the less-sensitive pipelines, even if they are only slightly less sensitive than the best one.

A more natural way to incorporate the efficiency information in the upper limit procedure is to orient the exclusion surfaces according to the measured efficiencies. For plane exclusion surfaces, the simplest choice is

$$\vec{k} = \vec{\epsilon}. \tag{34}$$

We term this choice the *efficiency weighted combination*, or EFF.

Heuristically, the efficiency weighted combination is an intelligent choice because it places the largest emphasis on the measurements made by the most sensitive combinations of pipelines. To see one of the desirable properties of this choice, consider a repeated experiment. In an ensemble of experiments, the expected number of detections by each pipeline combination i is

$$\langle\vec{n}\rangle = \lambda_{\rm true}\vec{\epsilon} + \vec{b}\,. \tag{35}$$

Suppose the observed number of events is $\vec{n}'$ in one experiment, and $\vec{n}''$ in a second. Which measurement should give the higher upper limit? If $(\vec{n}''-\vec{n}')\cdot\vec{\epsilon} > 0$, then the second measurement is consistent with a higher limit on λ. If $(\vec{n}''-\vec{n}')\cdot\vec{\epsilon} = 0$, then the two measurements imply the same upper limit on λ. The choice $\vec{k} = \vec{\epsilon}$ for the exclusion contours enforces these requirements.

3.5. Example: Rate Limit vs. Amplitude

Consider once more the case of two pipelines A and B. Let us suppose that the target signals are characterized by an amplitude ρ, and that the detection efficiencies of A and B separately, $E_A = \epsilon_A + \epsilon_{AB}$, $E_B = \epsilon_B + \epsilon_{AB}$, and their logical combinations $\epsilon_A, \epsilon_B, \epsilon_{AB}$, are as shown in Figure 1. This scenario is typical of searches for gravitational-wave bursts by LIGO and similar detectors [12, 16, 17, 18, 14]. Our objective is to set an upper limit on λ as a function of the signal amplitude ρ.

Since both E_A and $E_B \to 1$ at large ρ, $\epsilon_{AB} \to 1$ as well, while ϵ_A and ϵ_B are nonzero for only a limited range of signal amplitudes. In this toy model, A is sensitive to slightly weaker signals than B, so $\epsilon_A > \epsilon_B$. However, since both ϵ_A and ϵ_B are nonzero, each pipeline is able to detect some signals that the other pipeline misses. Therefore, one expects that combining the measurements of the two pipelines should be able to provide more information on the event rate than either pipeline alone.

Let us now compare the performance of four different choices of exclusion surfaces: AND, OR, SINGLE, and EFF. For the moment, let us ignore any background when computing the upper limits; *i.e.,* we will use $\vec{b} = 0$. (We will compare limits including background in the next section.)

Consider first the case where no events are detected. The upper limits from each combination are shown in Figure 2. All combinations give $\lambda_{90\%} = 2.3$ at high amplitude, where $\epsilon_{AB} \to 1$. In particular, the EFF upper limit is

$$\begin{aligned} 0.1 &= C_{\vec{k}}((0,0,0)|\lambda_{90\%}\vec{\epsilon}) \\ &= \mathrm{e}^{-\epsilon_{TOT}\lambda_{90\%}}\,, \end{aligned} \tag{36}$$

$$\lambda_{90\%} = \frac{2.30}{\epsilon_{TOT}}\,, \tag{37}$$

identical to the OR limit. The EFF and OR combinations give the strongest limits for weak signals because of their better efficiency (which is ϵ_{TOT}, the sum of the efficiencies of all pipeline combinations).

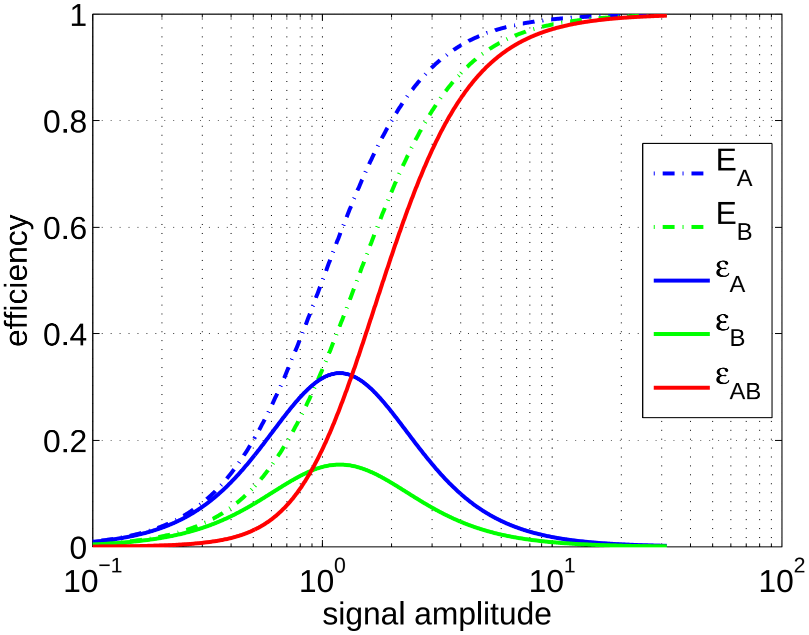

Figure 1. Efficiencies for two pipelines A and B. In our toy model, the signal is characterized by an amplitude ρ. The dotted lines $E_A = \epsilon_A + \epsilon_{AB}, E_B = \epsilon_B + \epsilon_{AB}$ show the efficiencies of the two pipelines considered separately. The continuous lines show the efficiencies of the logical combinations of the pipelines: ϵ_A (A not B), ϵ_B (B not A), and ϵ_{AB} (A and B).

Now consider the case of one event detected by the weaker pipeline B: $\vec{n} = (0, 1, 0)$. The upper limits are shown in Figure 3. The OR combination does poorly at high amplitudes because of the detected event. The AND limit is much better at high amplitudes because A did not see the event, but still poor at low amplitudes because $\epsilon_{AB} \to 0$. The SINGLE combination performs well, giving the same result as the $n = 0$ case, because it ignores the event counted by the less sensitive pipeline. The EFF upper limit is computed by summing over

$$\vec{N} \cdot \vec{\epsilon} \leq \vec{n} \cdot \vec{\epsilon} = \epsilon_B \, . \tag{38}$$

For signal amplitudes $\rho \gtrsim 1$, ϵ_B is the smallest efficiency, so the allowed terms are $\vec{N} \in \{(0, 0, 0), (0, 1, 0)\}$. The EFF limit is then given by

$$\begin{aligned} 0.1 &= P((0, 0, 0)|\lambda_{90\%}\vec{\epsilon}) + P((0, 1, 0)|\lambda_{90\%}\vec{\epsilon}) \\ &= (1 + \epsilon_B \lambda_{90\%}) \mathrm{e}^{-\epsilon_{TOT}\lambda_{90\%}} \, . \end{aligned} \tag{39}$$

The extra $\epsilon_B\lambda$ term makes the EFF upper limit only slightly higher than the $2.3/\epsilon_{TOT}$ value obtained in the $n = 0$ case, as can be seen from Figure 3. For $\rho \lesssim 1$, $\epsilon_B > \epsilon_{AB}$ and the upper limit includes additional $(\epsilon_{AB}\lambda)^{N_{AB}}$ terms. This causes the EFF limit to increase, but again only slightly, as ϵ_{AB} is typically much smaller than ϵ_A, ϵ_B at these low amplitudes.

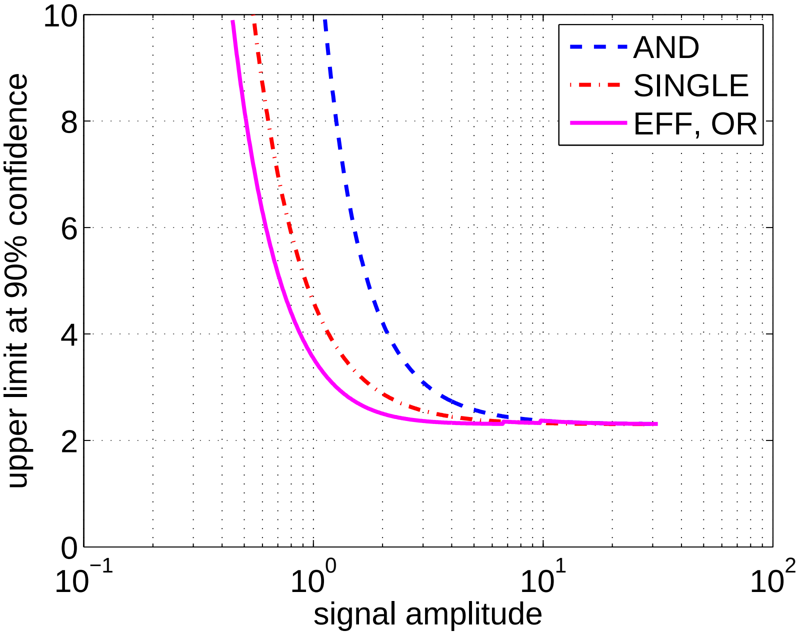

Figure 2. Upper limits as a function of signal amplitude when no events are detected. All methods give the asymptotic limit 2.3 for large amplitudes. The EFF and OR combinations give the strongest limits at low amplitude because they have better detection efficiency than the AND and SINGLE combinations.

We see that the EFF combination effectively ignores the event counted by the insensitive pipeline combination B, and gives a limit as good as or even slightly better than that from the SINGLE combination.

Now turn to the case in which one event is detected by the more sensitive pipeline, A: $\vec{n} = (1, 0, 0)$. The upper limits are shown in Figure 4. Again, the OR combination does poorly at high amplitudes because of the detected event. The SINGLE combination does even worse, since the event was found by the more sensitive pipeline, and the SINGLE combination has lower efficiency than the OR combination. The AND combination again performs well at high amplitudes and poorly at low amplitudes. The EFF upper limit is computed by summing over

$$\vec{N} \cdot \vec{\epsilon} \leq \vec{n} \cdot \vec{\epsilon} = \epsilon_A \,. \tag{40}$$

The number of terms in the sum depends on the relative values of ϵ_A, ϵ_B, and ϵ_{AB}. In this simple example, for $\rho > 1.4$, $\epsilon_A = 2\epsilon_B < \epsilon_{AB}$ and the allowed terms are $\vec{N} \in \{(0,0,0), (1,0,0), (0,1,0), (0,2,0)\}$. The EFF limit is given by

$$0.1 = (1 + \epsilon_A\lambda_{90\%} + \epsilon_B\lambda_{90\%} + {\epsilon_B}^2\lambda_{90\%}^2)\mathrm{e}^{-\epsilon_{TOT}\lambda_{90\%}} \,. \tag{41}$$

Since ϵ_A and ϵ_B are small at high amplitudes, the upper limit is again similar to the $n = 0$ value of $2.3/\epsilon_{TOT}$. For $\rho < 1.4$, $\epsilon_A > \epsilon_{AB}$ and the cumulative distribution $C_{\vec{k}}(\vec{n}|\lambda_\alpha\vec{\epsilon})$ in (19) includes additional $(\epsilon_{AB}\lambda)^{N_{AB}}$ terms. This causes the EFF limit to increase, becoming similar to that from the OR combination. In short, the EFF combination

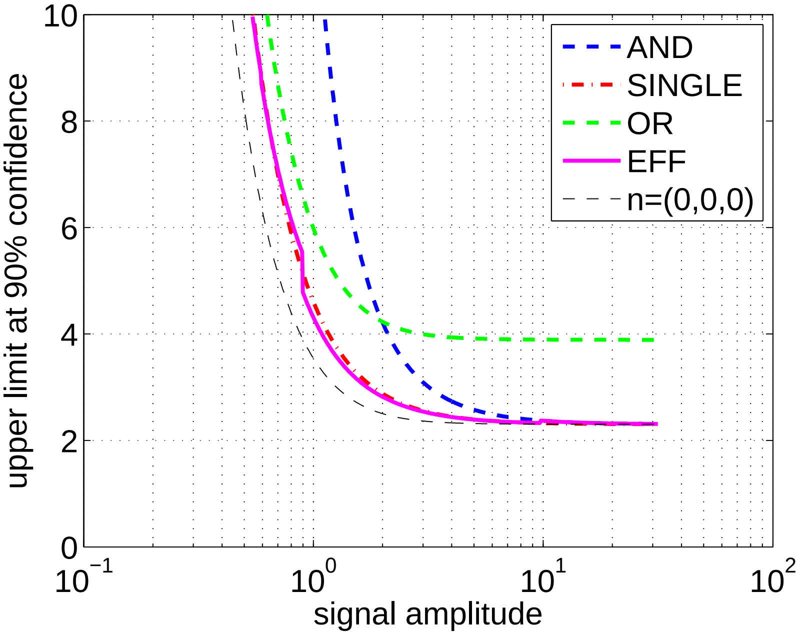

Figure 3. Upper limits as a function of signal amplitude when one event is detected by the less sensitive pipeline (B). The OR combination asymptotes to the single-event value 3.9. The lone event is not counted by the AND, SINGLE combinations, which give the $n = 0$ limit 2.3. The EFF combination ignores the event at high amplitudes (where $\epsilon_B \to 0$), while at lower amplitudes the EFF limit is very close to the SINGLE limit as $\epsilon_B \ll \epsilon_A$. The thin dashed line is the best possible upper limit from the counting experiment: that for zero observed events using the EFF or OR combinations (see Figure 2).

gives the strongest limits at high amplitudes because pipeline B should have seen a real event there and did not, and it gives the strongest limits at low amplitudes because it has better efficiency than the AND combination.

Finally, consider the case of a single event detected by both pipelines: $\vec{n} = (0, 0, 1)$. In this case all combinations give the asymptotic limit of 3.9 at large amplitudes, as seen in Figure 5. The relative limits of the AND, OR, and SINGLE combinations are the same as in the $n = 0$ case. We see, however, that the EFF combination outperforms all other combinations (including OR) in the low-amplitude limit. In fact, the EFF limit reaches nearly the $n = 0$ value at low signal amplitudes. This counter-intuitive result has a simple explanation: at low amplitudes ($\rho < 1$), the probability ϵ_{AB} of a real event being detected jointly by A and B is much smaller than the probabilities ϵ_A, ϵ_B of it being detected by either pipeline alone. The observation $n_{AB} > 0$ is therefore inconsistent with the hypothesis of a low-amplitude signal. The efficiency weighted combination therefore ignores this measurement for the low-amplitude upper limits, and the limit is dominated by the measurements $n_A = 0 = n_B$.

It is worth noting that the upper limits obtained from the efficiency-weighted procedure are neither monotonic nor continuous; this is most evident in Figure 5. The

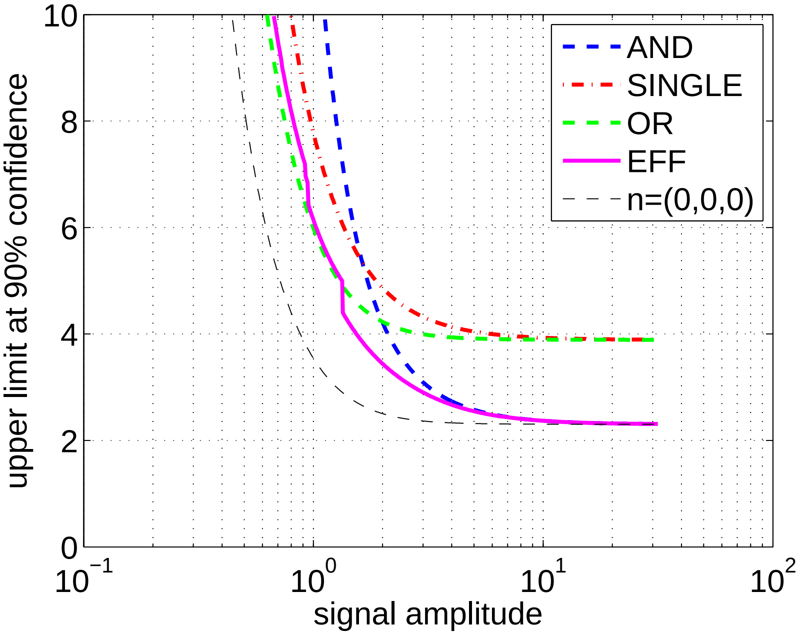

Figure 4. Upper limits as a function of signal amplitude when one event is detected by the more sensitive pipeline (A). The OR and SINGLE combinations asymptote to the single-event value 3.9. The lone event is not counted by the AND combination, which gives the $n = 0$ limit 2.3. The EFF combination ignores the event at high amplitudes (where $\epsilon_A \to 0$), while at lower amplitudes the EFF limit is very close to the OR limit for $n = 1$. The thin dashed line is the best possible upper limit from the counting experiment: that for zero observed events using the EFF or OR combinations (see Figure 2).

limits are not monotonic because the efficiencies ϵ_A, ϵ_B, ϵ_{AB} of the logical combinations of pipelines are not monotonic, as shown in Figure 1. The origin of the discontinuities is slightly more subtle; it arises from the need to sum over a discrete set of $\vec{N}$ in (19). For the efficiency-weighted combination, the condition (20) depends on the assumed signal amplitude through the efficiencies, $\vec{k} = \vec{\epsilon}(\rho)$. Therefore, the sum may include different numbers of terms for different signal amplitudes. The discontinuities occur at signal amplitudes where another term satisfies the condition to be included in the sum in (19), $\vec{N} \cdot \vec{\epsilon} \leq \vec{n} \cdot \vec{\epsilon}$. In turn, this happens when the ratio of efficiencies equals a rational number. We stress that these discontinuities are a general feature of using efficiencies to weight the pipeline combinations, and that they are not indicative of any problem with the procedure. The upper limits at different ρ values are limits on different signal models, and therefore they need not be continuous or monotonic functions of ρ. Indeed, this behaviour is advantageous, as seen in Figure 5, where the efficiency-weighted upper limit is able to drop below the OR limit at low amplitudes.

In each of the cases considered, efficiency weighting gives upper limits as approximately as strong as or stronger than any of the other choices. Without efficiency weighting, the best remaining combination is different for the different cases: AND,

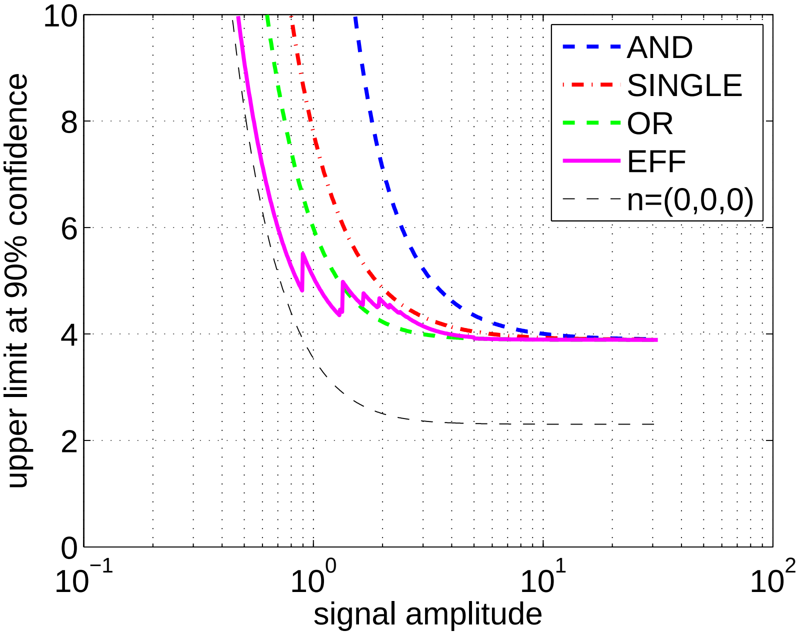

Figure 5. Upper limits as a function of signal amplitude when a single event is detected by both pipelines (A and B). All combinations give the asymptotic limit of 3.9 at large amplitudes. The EFF combination ignores this event at low amplitudes (where $\epsilon_{AB} \ll \epsilon_A, \epsilon_B$) and tends to the zero-event limit for $\rho < 1$. The thin dashed line is the best possible upper limit from the counting experiment: that for zero observed events using the EFF or OR combinations (see Figure 2).

OR, and SINGLE each perform best for at least one of the cases tested. While we must chose the weighting *before* measuring $\vec{n}$ for the upper limit procedure to have the proper coverage, there is no way to know *a priori* whether to choose AND, OR, or SINGLE. The efficiency-weighted combination, however, gives optimal or near-optimal performance in all cases.

We can gain insight into the strong performance of the efficiency weighting choice by examining the form of the upper limit equation (19):

$$1 - \alpha = (1 + \epsilon_1\lambda_\alpha + \frac{\epsilon_1^2\lambda_\alpha^2}{2} + \ldots + \epsilon_2\lambda_\alpha + \ldots)\mathrm{e}^{-(\epsilon_1+\ldots)\lambda_\alpha} \,. \qquad (42)$$

The set of efficiencies ϵ_i appearing in the exponential is determined by the choice of pipeline combination used for the upper limit. The set of ϵ_i terms appearing in the factor in front of the exponential depends on the set of measured events $\vec{n}$ as well as the pipeline combination chosen. As a rule, adding efficiency terms in the exponential *decreases* the upper limit. Adding efficiency terms to the factor in front of the exponential *increases* the upper limit. For the AND and SINGLE combinations, only some of the efficiencies appear in the exponential. With the OR and EFF combinations, the efficiencies for all pipeline combinations appear in the exponential, giving the maximum efficiency possible (ϵ_{TOT}). Between these two, the EFF combination will typically give fewer terms in the prefactor when events are detected with the less sensitive pipeline combinations. This will result in a lower limit than the OR combination. It may have more terms when the most sensitive combination sees the event, thus giving a higher limit than the OR combination in these cases. As seen in Figure 5, this loss in upper limit tends to be small; since the extra terms are associated with low-efficiency pipeline combinations, and appear with powers of those small ϵ_i.

3.6. Upper Limits with Background

We have seen that the EFF weighted combination tends to give stronger upper limits than the AND, OR, and SINGLE weightings when we ignore the background. We now demonstrate by example that this superior performance continues when we account for the background as well. We do this by computing the expectation value of the upper limit as a function of the true foreground rate λ for two scenarios: one with low background, and one with high background.

Let us consider once more the case of our two pipelines A and B. We will work initially with a fixed set of efficiencies,

$$\vec{\epsilon} = (\epsilon_A, \epsilon_B, \epsilon_{AB}) = (0.345, 0.175, 0.480)\,. \tag{43}$$

Let us assume the background to be

$$\vec{b} = (b_A, b_B, b_{AB}) = (1/3, 1/3, 1/3)\, b_{TOT}\,. \tag{44}$$

With this background, on average, pipelines A and B detect the same number of background events, and half of the events detected by one are also detected by the other. The total expected background is b_{TOT}. We will consider the cases $b_{TOT} = 0.1$ ("low background") and $b_{TOT} = 1$ ("high background").

A straightforward Monte Carlo analysis was used to estimate the upper limit in an ensemble of experiments. Figure 6 shows the mean limits from the AND, OR, SINGLE, and EFF combinations as a function of the true value of $\lambda \in [0, 1]$ for the low background case. Figure 7 shows the mean limits for the high background case. In both cases the EFF weighting gives stronger limits than any of the other weightings for all values of λ tested. The gap between the EFF upper limits and the next best limits (from OR) is particularly large for the high-background case. These findings support our conclusion that the EFF weighting "protects" the upper limit against modest background contamination.

To get a sense of the robustness of the EFF weighting performance, we repeat the Monte Carlo for a range of efficiencies. Specifically, we vary ϵ_A over [0,1], $\epsilon_B \leq \epsilon_A$, and keep $\epsilon_{AB} = 1 - \epsilon_A - \epsilon_B$ so that $\epsilon_{TOT} = 1$. We use $b_{TOT} = 1$ ("high background") and $\lambda = 0.5$. Figure 8 shows how the mean upper limit from the EFF weighting varies with ϵ_A, ϵ_B. The mean limits range from 2.91 to 3.41, a variation of less than 20%. By contrast, the mean limits from the other weightings (not shown) are always higher: $\geq$ 3.40 (SINGLE); $\geq$ 3.25 (AND), and = 3.55 (OR). This indicates that the superior

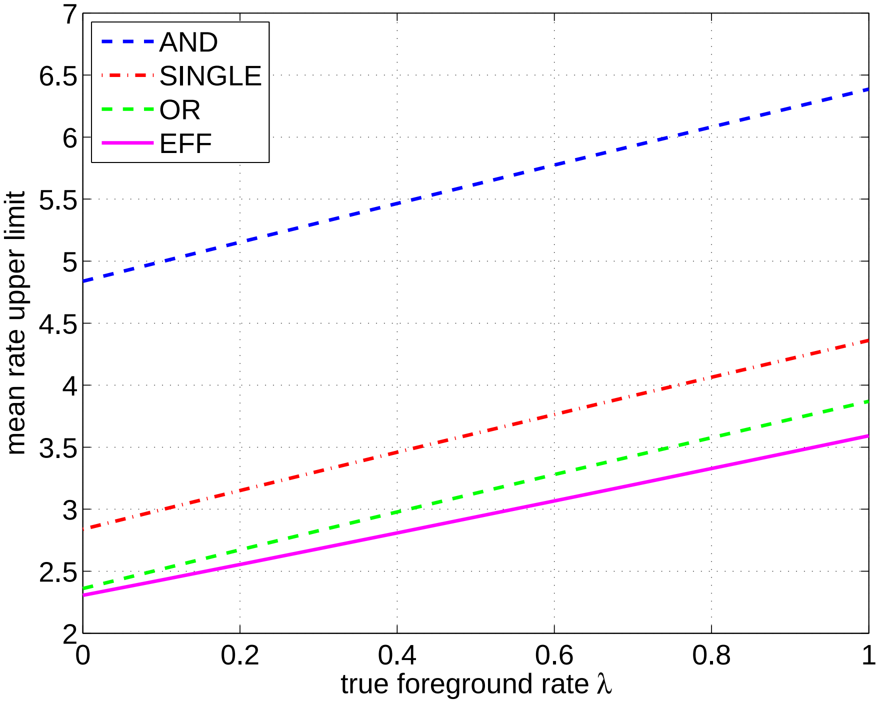

Figure 6. Mean upper limit as a function of the true foreground rate λ in an ensemble of experiments with fixed low background. This two-pipeline experiment has efficiency $(\epsilon_A, \epsilon_B, \epsilon_{AB}) = (0.345, 0.175, 0.480)$ and background $(b_A, b_B, b_{AB}) = (1/30, 1/30, 1/30)$.

performance of the EFF weighting is not reliant on the efficiencies taking particular values.

It can be noted from Figure 8 that the EFF limit does not reduce to the SINGLE limit (≥ 3.40) when $\epsilon_B \to 0$. This is because the EFF combination becomes $\vec{k} = (\epsilon_A, 0, \epsilon_{AB})$, whereas the SINGLE weighting is $\vec{k} = (1, 0, 1)$. So, the EFF weighting maintains a distinction between events detected by A alone and those detected jointly by A and B. The result is that the EFF limits are lower than or equal to the SINGLE limits as $\epsilon_B \to 0$, with equality at $\vec{\epsilon} = (0.5, 0, 0.5)$.

Finally we note that the EFF weighting, since it is based on efficiency alone, is most applicable to the case where the background is relatively small. We concentrate on the case where the expected number of events due to background of order 1 or less. For much higher backgrounds the optimal weightings should also include information on the backgrounds b_A, b_{AB}, . . . of the various pipeline combinations.

4. Multiple Data Sets

The formalism we have developed for multiple algorithms analyzing a common data set can be applied equally well to the analysis of multiple sets of data. For example, we may have data from several observation periods, each characterized by the use of a different

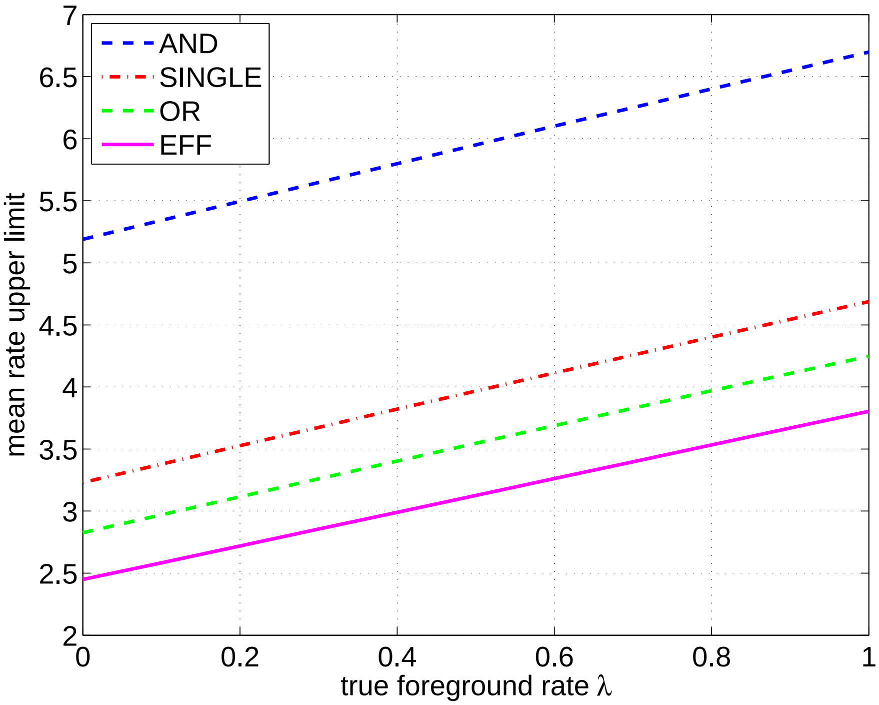

Figure 7. Mean upper limit as a function of the true foreground rate λ in an ensemble of experiments with fixed high background. This two-pipeline experiment has efficiency $(\epsilon_A, \epsilon_B, \epsilon_{AB}) = (0.345, 0.175, 0.480)$ and background $(b_A, b_B, b_{AB}) = (1/3, 1/3, 1/3)$.

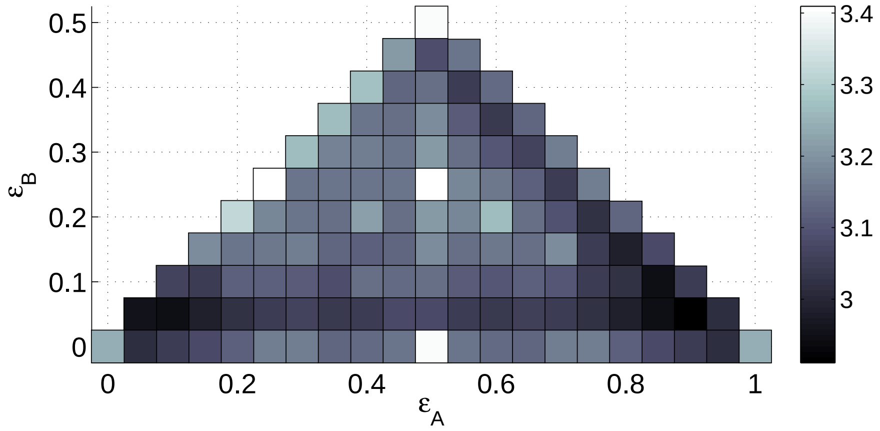

Figure 8. Mean upper limit as a function of efficiency $\vec{\epsilon} = (\epsilon_A, \epsilon_B, 1 - \epsilon_A - \epsilon_B)$ in an ensemble of experiments with background $(b_A, b_B, b_{AB}) = (1/3, 1/3, 1/3)$ and true event rate $\lambda = 0.5$. The largest limits occur when ϵ_A, ϵ_B, and $\epsilon_{AB} = 1 - \epsilon_A - \epsilon_B$ are related by the ratio of small integers, as discussed in Section 3.5.

set of instruments, or over which the sensitivity of the instruments changed, *etc.* In this case, the analyses of the separate data epochs may be considered as separate pipelines for purposes of setting an upper limit.

As a simple example, consider the case of a single algorithm used to analyse data from two disjoint data sets A and B, with durations T_A, T_B. The sensitivity of the experiment is characterized by the two numbers

ϵ_A: The probability that any given foreground event will be detected during period A;

ϵ_B: The probability that any given foreground event will be detected during period B.

The background is characterized by

b_A: The expected number of background events detected during period A;

b_B: The expected number of background events detected during period B.

The outcome of the experiment is the set of two numbers

n_A: The number of events detected during period A;

n_B: The number of events detected during period B.

Since any given event can be detected during period A or period B but not both, we have $\epsilon_{AB} = 0$, $b_{AB} = 0$, $n_{AB} = 0$. We see immediately that this is a special case of the two-pipeline analysis, where we treat the analysis of the separate data sets as separate pipeline measurements. In fact, it is a particularly simple case, as we know $\epsilon_{AB} = 0$, $b_{AB} = 0$, $n_{AB} = 0$ *a priori.*

Note that we define the efficiencies ϵ_A, ϵ_B in terms of the probability of events from anywhere in the *entire* observation period T being detected during periods A or B. We are taking the union of the data sets to treat them as one large set. This is the most convenient approach, since it matches precisely how the multiple-pipeline case was developed. It saves us from including the separate observation times T_A, T_B explicitly in our upper limit calculations. Instead, they are included implicitly in the efficiencies. For example, ϵ_A has a maximum possible value of $T_A/(T_A + T_B)$.

For concreteness, let us suppose we have two data sets of equal length, $T_A = T_B = 0.5T$. Suppose also that the instruments used were more sensitive during period A, such that $\epsilon_A = 3/5$, $\epsilon_B = 2/5$, and $\epsilon_{TOT} = \epsilon_A + \epsilon_B = 1$. Table 1 shows the upper limits obtained ignoring the background. We compare the OR (combining event counts from both periods), SINGLE (only counting events from the more sensitive period), and EFF combinations for zero or one detected event. (The AND combination is not applicable to this case, since $\epsilon_{AB} = 0$.)

In each case, the EFF combination gives the best upper limit. For no detected events, the EFF and OR combinations give the limit 2.3 as before. The SINGLE limit is a factor 5/3 higher, because it uses only 3/5 of the integrated sensitivity of the experiment ($\epsilon_A = 3\epsilon_{TOT}/5$). For one event detected in the less sensitive period B, the EFF combination gives the best upper limit – even better than SINGLE. This may be surprising, in that the SINGLE upper limit is computed for zero events. We see that

Table 1. Comparison of upper limits obtained for various possible outcomes of a counting experiment on two data sets A and B with $\epsilon_A = 3/5$, $\epsilon_B = 2/5$, and ignoring background. The cases are: no events detected ($\vec{n} = (0,0,0)$); one event detected in B ($\vec{n} = (0,1,0)$); one event detected in A ($\vec{n} = (1,0,0)$).

	upper limit		
$\vec{n}$	OR	SINGLE	EFF
(0,0,0)	2.3	3.8	2.3
(0,1,0)	3.9	3.8	3.1
(1,0,0)	3.9	6.5	3.9

the extra sensitivity gained by including the B measurement in the EFF upper limit more than offsets the loss in the limit due to having a detected event. Finally, for the case of one event detected in the more sensitive period A, the EFF limit matches the OR limit. Interestingly enough, the SINGLE combination performs worse than EFF in all cases; for the given efficiencies, we always get a better limit by using all of the data.

For a larger difference in efficiencies, the differences in upper limits are more pronounced. Table 2 compares the upper limits for $\epsilon_A = 2/3$, $\epsilon_B = 1/3$, $\epsilon_{TOT} = 1$. The OR limits are unchanged. The SINGLE limits are better than those in Table 1 because the SINGLE combination now contains 2/3 of the integrated sensitivity of the experiment ($\epsilon_A = 2\epsilon_{TOT}/3$) instead of only 3/5. The changes in the EFF limits are more complicated. For one event detected in the less sensitive period B, the EFF combination still gives the best upper limit – slightly better than before, because the weighting of B is less than in the previous case. For one event detected in A, the EFF limit is between the OR limit and the SINGLE limit. The increase over the limit in Table 1 is due to the fact that for $\epsilon_A = 2\epsilon_B$, the cumulative sum in (19) now includes the terms $\vec{N} = \{(0,0,0),(1,0,0),(0,1,0),(0,2,0)\}$, whereas for $\epsilon_A = 1.5\epsilon_B$ it includes only $\vec{N} = \{(0,0,0),(1,0,0),(0,1,0)\}$.

Note that the EFF limits are particularly robust against background events contaminating the less-sensitive data sets. This allows the sub-optimal data to be used to strengthen scientific results without fear of "spoiling" the upper limits. In particular, note that the average of the upper limits for the single-event cases ($\vec{n} = (1,0,0)$ and $(0,1,0)$) is best for the EFF combination in both Table 1 and Table 2. So, if the data sets have equal background probability (assumed $\ll 1$), the EFF combination will on average give the best upper limits for low true event rates.

Finally, since we have seen benefits from treating multiple data sets separately, one might ask if we should always split up data sets. In particular, why not sub-divide all data sets *ad infinitum*? The answer comes from noting that the benefits of the EFF combination arise from exploiting differences in the efficiencies ϵ_i. If the differences in efficiency between two data sets are negligible, then there is no benefit to treating them separately. For example, for two sets of data with identical efficiencies, the EFF and OR combinations will always give identical limits: since $n_{AB} = 0$ always, choosing $\vec{k} = \vec{\epsilon}$

Table 2. Comparison of upper limits obtained for various possible outcomes of a counting experiment on two data sets A and B with $\epsilon_A = 2/3$, $\epsilon_B = 1/3$, and ignoring background. The cases are: no events detected ($\vec{n} = (0,0,0)$); one event detected in B ($\vec{n} = (0,1,0)$); one event detected in A ($\vec{n} = (1,0,0)$).

	upper limit		
$\vec{n}$	OR	SINGLE	EFF
(0,0,0)	2.3	3.5	2.3
(0,1,0)	3.9	3.5	3.0
(1,0,0)	3.9	5.8	4.3

will always give the same results as $\vec{k} = (1, \ldots, 1)$. One therefore gets no benefit from sub-dividing epochs of constant sensitivity.

5. Summary

We have proposed a general technique for setting upper limits on Poisson processes from counting experiments involving multiple data sets and multiple event-counting algorithms (which we collectively refer to as multiple "pipelines"). This technique is an extension of the standard procedure for one-sided classical confidence intervals. There are two key features. First, we characterize the measurements by the logical combinations of pipelines – the number of events counted by A-and-B, by A-and-not-B, *etc.* Second, we select a rank-ordering of the space of possible measurements which is based on the relative detection efficiencies of these logical combinations. This efficiency weighting uses all of the counts from the experiment, but assigns more significance to those counts from pipeline combinations which are expected to detect more foreground events. We have seen that in typical cases for low background and low foreground event rate, the efficiency weighting tends to give stronger upper limits than selecting the AND or OR combination of pipelines, or selecting the single most sensitive pipeline only. In particular, the efficiency weighting procedure tends to be robust against modest background contamination of the event counts. This allows all of the observational results to contribute to the upper limit while reducing the chances that background contamination of some counts will weaken it.

In this paper we have focused on computing upper limits; however, the method has wider applicability. The characterisation of the experiment in terms of logical combinations of pipelines and the subsequent rank ordering effectively reduce the space of measurements to one dimension. At this point we are free to apply other standard procedures for constructing one- or two-sided confidence intervals. It would be interesting, for example, to apply the Feldman-Cousins procedure [1] to produce unified upper limits and confidence intervals for our multiple-pipeline experiment; we leave this consideration to the future.

As a final note, let us point out that the concept of a "pipeline" is quite general –

it is nothing more than a way of defining a count of events. We have seen that different pipelines may consist of different algorithms applied to the same data, or the same algorithm applied to different data sets. Distinct pipelines may also be defined in other ways, such as by applying a single algorithm to a single data set and segregating the resulting events into groups by some other attribute. For example, in gravitational-wave burst searches the background is largely due to events detected at low frequencies (< 200 Hz). Dividing events into low-frequency (< 200 Hz) and high-frequency (> 200 Hz) sets would produce limits on high-frequency gravitational waves that are not compromised by the low-frequency background. LIGO matched-filtering searches for gravitational waves from inspiralling binaries [19, 20] use a similar idea, dividing the space of templates (signal parameters) into several regions. Background events that match templates in one region then have minimal impact on the limits set in other regions of the template/signal space. Multiple applications of the same algorithm with different counting thresholds can also be treated as separate pipelines and handled by our method; this might be appropriate when low- and high-amplitude events are produced by separate populations. A multiple-threshold approach would have the benefit that events detected with low (high) amplitude have minimal impact on the rate limits set on the high (low) amplitude population. Our method even naturally handles the case of a "veto" analysis, in which one pipeline (B) processes data in such a way as to be deliberately *insensitive* to signals, but sensitive to background noise: ϵ_B, $\epsilon_{AB} \simeq 0$ but b_B, $b_{AB} > 0$. The EFF weighting then automatically ignores (vetoes) events detected by A that are also detected by the veto pipeline B.

Acknowlegements

The author would like to thank Lindy Blackburn, Shourov Chatterji, Jolien Creighton, Stephen Fairhurst, and Erik Katsavounidis for stimulating discussions, and Patrick Brady for valuable feedback on an earlier draft. This work was supported in part by STFC grant PP/F001096/1.

Appendix A. Derivative of $C_{\mathcal{S}}(\vec{n}|\vec{\epsilon},\lambda)$

In this appendix we prove equation (18),

$$\frac{\mathrm{d}C_{\mathcal{S}}(\vec{n}|\lambda\vec{\epsilon}+\vec{b})}{\mathrm{d}\lambda} < 0\,, \tag{A.1}$$

where $\vec{n}$, $\vec{\epsilon}$, $\vec{b}$, and the family $\mathcal{S}(\zeta)$ are held fixed.

First, we recall the definition (15) of the cumulative probability $C_{\mathcal{S}}$,

$$\begin{aligned} C_{\mathcal{S}}(\vec{n}|\lambda\vec{\epsilon}+\vec{b}) &= \sum_{\vec{N}|\zeta(\vec{N})\leq\zeta(\vec{n})} P(\vec{N}|\lambda\vec{\epsilon}+\vec{b}) \\ &= \sum_{\vec{N}|\zeta(\vec{N})\leq\zeta(\vec{n})} \prod_{i=1}^{q} P(N_i|\epsilon_i\lambda+b_i) \end{aligned}$$

$$= \sum_{\vec{N}|\zeta(\vec{N})\leq\zeta(\vec{n})} \prod_{i=1}^{q} \frac{(\epsilon_i\lambda + b_i)^{N_i} \mathrm{e}^{-\epsilon_i\lambda - b_i}}{N_i!} \,. \tag{A.2}$$

Taking the derivative with respect to λ yields

$$\begin{aligned}\frac{\mathrm{d}C_{\mathcal{S}}(\vec{n}|\lambda\vec{\epsilon}+\vec{b})}{\mathrm{d}\lambda} = & \sum_{\vec{N}|\zeta(\vec{N})\leq\zeta(\vec{n})} \left[\frac{N_1\epsilon_1(\epsilon_1\lambda+b_1)^{N_1-1}}{N_1!} \times \ldots \times \frac{(\epsilon_q\lambda+b_q)^{N_q}}{N_q!} + \ldots \right. \\ & + \frac{(\epsilon_1\lambda+b_1)^{N_1}}{N_1!} \times \ldots \times \frac{N_q\epsilon_q(\epsilon_q\lambda+b_q)^{N_q-1}}{N_q!} \\ & \left. - (\epsilon_1 + \ldots + \epsilon_q) \frac{(\epsilon_1\lambda+b_1)^{N_1}}{N_1!} \times \ldots \times \frac{(\epsilon_q\lambda+b_q)^{N_q}}{N_q!} \right] \\ & \times \exp\left[-\left(\epsilon_1 + \ldots + \epsilon_q\right)\lambda - \left(b_1 + \ldots + b_q\right)\right] \,. \end{aligned} \tag{A.3}$$

Consider the contribution of the term $\vec{N}' = (N'_1, \ldots, N'_q)$ to the sum. We see that the positive terms arise from taking the derivative of the λ^{N_i}. Each such positive term is exactly cancelled by a negative term coming from the derivative of the exponential from the term $\vec{N}'' = (N'_1, \ldots, N'_i - 1, \ldots, N'_q)$. $\vec{N}''$ will always be included in the sum if $\vec{N}'$ is included because of the requirement that the normal to the surfaces $\mathcal{S}(\zeta)$ must have only non-negative components. Therefore, all positive terms in (A.3) are cancelled and the derivative must be negative.

References

[1] Gary J. Feldman and Robert D. Cousins. Unified approach to the classical statistical analysis of small signals. *Phys. Rev. D*, 57:3873–3889, 1998.

[2] Patrick R. Brady, Jolien D. E. Creighton, and Alan G. Wiseman. Upper limits on gravitational-wave signals based on loudest events. *Class. Quant. Grav.*, 21:S1775–S1782, 2004.

[3] Rahul Biswas, Patrick R. Brady, Jolien D. E. Creighton, and Stephen Fairhurst. The loudest event statistic: general formulation, properties and applications. *Classical and Quantum Gravity*, 26:175009, 2009.

[4] O. Helene. Errors in experiments with small numbers of events. *Nuclear Instruments and Methods in Physics Research A*, 228:120–128, 1984.

[5] H. B. Prosper. A Bayesian analysis of experiments with small numbers of events. *Nuclear Instruments and Methods in Physics Research A*, 241:236–240, 1985.

[6] Harrison Bertrand Prosper. Small-signal analysis in high-energy physics: A bayesian approach. *Phys. Rev. D*, 37:1153–1160, 1988.

[7] D. A. Williams. Comment on "small-signal analysis in high-energy physics: A bayesian approach". *Phys. Rev. D*, 38:3582–3583, 1988.

[8] Harrison Bertrand Prosper. Reply to "comment on 'small-signal analysis in high-energy physics: A bayesian approach' ". *Phys. Rev. D*, 38:3584–3585, 1988.

[9] `http://www.ligo.caltech.edu/`.

[10] `http://www.virgo.infn.it/`.

[11] `http://www.geo600.uni-hannover.de/`.

[12] B. Abbott et al. First upper limits from ligo on gravitational wave bursts. *Phys. Rev. D*, 69:102001, 2004.

[13] F Beauville et al. A first comparison of search methods for gravitational wave bursts using ligo and virgo simulated data. *Classical and Quantum Gravity*, 22:S1293–S1301, 2005.

[14] B. P. Abbott et al. Search for gravitational-wave bursts in the first year of the fifth LIGO science run. 2009. arXiv:0905.0020.

[15] B P Abbott et al. Ligo: the laser interferometer gravitational-wave observatory. *Reports on Progress in Physics*, 72, 2009.

[16] B. Abbott et al. Upper limits on gravitational wave bursts in ligo's second science run. *Phys. Rev. D*, 72:062001, 2005.

[17] B. Abbott et al. Upper limits from the ligo and tama detectors on the rate of gravitational-wave bursts. *Phys. Rev. D*, 72:122004, 2005.

[18] B. Abbott et al. Search for gravitational-wave bursts in ligo data from the fourth science run. *Class. Quantum Grav.*, 24:5343–5369, 2007.

[19] B. P. Abbott et al. Search for gravitational waves from low mass binary coalescences in the first year of ligo's s5 data. *Physical Review D*, 79:122001, 2009.

[20] B. P. Abbott et al. Search for gravitational waves from low mass compact binary coalescence in 186 days of ligo's fifth science run. *Physical Review D*, 80:047101, 2009.